The Quest for Dense Structure in Heterogeneous Data Exploration

Charles Phillips
Department of Electrical Engineering and Computer Science
University of Tennessee

CiSML Seminar
3/27/2015
Outline of Talk

• Heterogeneous Data
• Accounting for Noise
• Synthetic Data
• Sanity Checking
• Similarity/Distance Metrics
• Thresholding
• Dense Structure
• Problem Equivalence
• Hypothesis Generation/Data Exploration
• Importance of Domain Knowledge
Heterogeneous Data

- Genes
- Phenotypes
- People
- Articles
- Customers
- Authors
- Products
- Movies
- Proteins
- Diseases
Sources of Noise

• Instrument precision
• Instrument malfunctions
• User malfunctions
• Missing values
  – Not the same as zero!
• Confounds
Noise Examples

• Microarray fingerprint
• PCA Analysis on dozens of phenotypes
• Selecting for traits in mice
• Low dose radiation

Synthetic Data

- Random Data
- Simulated Data
  - Is signal modeled correctly?
  - Is noise modeled correctly?
- Ground truth is unknown
- Differences between synthetic and real biological data
  - Degree distribution
  - Clique size
Sanity Checking

$n = 62$

$n = 4$
Sanity Checking

![Image of a spreadsheet showing data analysis with columns for streetname, apt_no, city, stat, zip6, zip4, x_street, num_unit, res_inv, acres_burn, less_on_site, map, area_ign, conf_orig, and fact_ign_2.](image)
## Similarity/Distance Metrics

<table>
<thead>
<tr>
<th>Metric</th>
<th>Measures</th>
<th>Notes/Uses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation</td>
<td>Linear dependence</td>
<td>Many</td>
</tr>
<tr>
<td>Spearman Correlation</td>
<td>Linear dependence of ranks</td>
<td>Resistant to outliers</td>
</tr>
<tr>
<td>Mutual Information</td>
<td>Non-linear dependence</td>
<td>Categorical variables</td>
</tr>
<tr>
<td>Cosine Similarity</td>
<td>Similarity in vector space</td>
<td>Document comparison</td>
</tr>
<tr>
<td>Euclidean Distance</td>
<td>Distance in Cartesian space</td>
<td>Straight line distance</td>
</tr>
<tr>
<td>Orthodromic Distance</td>
<td>Distance on a sphere</td>
<td>Geocoding (simple method)</td>
</tr>
<tr>
<td>Jaccard Similarity</td>
<td>Similarity between sets</td>
<td>Set-set comparison</td>
</tr>
<tr>
<td>Hamming Distance</td>
<td>Strings of equal length</td>
<td>Error correction</td>
</tr>
<tr>
<td>Levenshtein Distance</td>
<td>Strings, where insertions and deletions are expected</td>
<td>DNA sequence comparison</td>
</tr>
<tr>
<td>Known Association</td>
<td>Molecular interaction, friends, purchases</td>
<td>Protein-protein interactions, social networks, customers</td>
</tr>
</tbody>
</table>
Thresholding

- Similarity Matrix
- Threshold Selection
  - Spectral Methods
  - Analogous to p-value selection
  - Tradeoff between false positives and false negatives
- Unweighted Graphs
Thresholding

- 37 recombinant inbred BXD mouse strains
- Case group: ethanol, Control group: saline
- Differential expression between case and control (S-scores)

22626 Genes
2137 Quantitative Phenotypes
Thresholding

Phenotypes and mRNA probes exhibit different correlation profiles. There are fewer inter-type correlations than intra-type correlations.
## Dense Structure

<table>
<thead>
<tr>
<th>Paraclique</th>
<th>Genes</th>
<th>Phenotypes</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>9</td>
<td>68</td>
<td>All but 4 phenotypes are cocaine response</td>
</tr>
<tr>
<td>6</td>
<td>26</td>
<td>55</td>
<td>All phenotypes are morphine response</td>
</tr>
<tr>
<td>26</td>
<td>5</td>
<td>21</td>
<td>All phenotypes are morphine response</td>
</tr>
<tr>
<td>28</td>
<td>8</td>
<td>15</td>
<td>All phenotypes are ethanol response</td>
</tr>
</tbody>
</table>
Dense Structure

• **Clique**
  – A set of vertices with all possible edges.
  – A complete subgraph.

• **Maximum Clique**
  – The largest clique in a graph

• **Maximal Clique**
  – A clique to which no vertex can be added to form a larger clique.
  – A clique that is not a proper subset of another clique.

![Graph diagram with vertices and edges labeled a, b, c, d, e, f, g, h.]
Dense Structure

• **Clique**
  – A set of vertices with all possible edges.
  – A complete subgraph.

• **Maximum Clique**
  – The largest clique in a graph

• **Maximal Clique**
  – A clique to which no vertex can be added to form a larger clique.
  – A clique that is not a proper subset of another clique.

- Maximum clique: \{ a, b, d, e \}
- Maximal cliques: \{ a, b, d, e \}
  \{ c, f, g \}
  \{ c, g, h \}
  \{ e, f \}
Dense Structure

• $k$-clique
  – A clique with $k$ vertices.

• Biclique
  – A complete bipartite graph
  – $K_{m,n}$
  – All possible interpartite edges

• Maximal Biclique
  – A biclique to which no vertex can be added to form a larger biclique
  – A biclique that is not a superset of any other biclique
Results

The graph shows the log of the wallclock runtime in seconds for different p-value thresholds. The runtime increases as the p-value threshold decreases. The graph compares three methods: MBEA, LCM, and MICA.
Vertex and Edge Maximum

Bipartite Graph

Vertex-Maximum Biclique
(8 vertices, 7 edges)
Polynomial Time

Edge-Maximum Biclique
(6 vertices, 9 edges)
NP-hard

Tripartite Graph

Vertex-Maximum Triclique
(7 vertices, 10 edges)
NP-hard

Edge-Maximum Triclique
(6 vertices, 12 edges)
NP-hard
Transaction Data
Transaction Data

• Also known as *market basket* data. Each transaction is a set of items.
  {Bread, Milk, Cheese}
  {Milk, Cheese, Eggs, Sugar}
  {Bread, Milk, Eggs}
  {Cheese, Eggs, Sugar, Flour}

• *Itemset*
  — A set of items.

• *Support*
  — The number of transactions in which an itemset occurs.

• *Frequent Itemset*
  — An itemset whose support is at or above some specified threshold.

• *Closed Itemset*
  — An itemset that has no superset with the same support.

• *Maximal Itemset*
  — An itemset that has no superset that is frequent.
## Problem Equivalence

<table>
<thead>
<tr>
<th>Transaction</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ABDEF</td>
</tr>
<tr>
<td>2</td>
<td>ABF</td>
</tr>
<tr>
<td>3</td>
<td>BCDE</td>
</tr>
<tr>
<td>4</td>
<td>ABCE</td>
</tr>
<tr>
<td>5</td>
<td>CDE</td>
</tr>
<tr>
<td>6</td>
<td>CDEF</td>
</tr>
<tr>
<td>7</td>
<td>AEF</td>
</tr>
</tbody>
</table>
Problem Equivalence

<table>
<thead>
<tr>
<th>Transaction</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ABDEF</td>
</tr>
<tr>
<td>2</td>
<td>ABF</td>
</tr>
<tr>
<td>3</td>
<td>BCDE</td>
</tr>
<tr>
<td>4</td>
<td>ABCE</td>
</tr>
<tr>
<td>5</td>
<td>CDE</td>
</tr>
<tr>
<td>6</td>
<td>CDEF</td>
</tr>
<tr>
<td>7</td>
<td>AEF</td>
</tr>
</tbody>
</table>

A
B
C
D
E
F
Problem Equivalence

<table>
<thead>
<tr>
<th>Transaction</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ABDEF</td>
</tr>
<tr>
<td>2</td>
<td>ABF</td>
</tr>
<tr>
<td>3</td>
<td>BCDE</td>
</tr>
<tr>
<td>4</td>
<td>ABCE</td>
</tr>
<tr>
<td>5</td>
<td>CDE</td>
</tr>
<tr>
<td>6</td>
<td>CDEF</td>
</tr>
<tr>
<td>7</td>
<td>AEF</td>
</tr>
</tbody>
</table>

Frequent Itemset \((s=1)\) = Biclique
Closed Itemset = Maximal Biclique
Closed Frequent Itemset = Maximal Biclique with \(|U| > s\)
The red vertices are a maximal biclique in partite sets $A$ and $B$. 
The red vertices are a maximal triclique in the graph. But the red vertices in $A$ and $B$ are not a maximal biclique.
Dense Structure

The red vertices in the lefthand graph form a maximum biclique between partite sets A and B. But a maximum triclique in the same graph on the right only contains a subset of the maximum biclique.
Tripartite

18 Allen Brain Atlas Genesets

50 Genesets Returned by GeneWeaver Query “Alcohol + Ethanol”

10821 Mammalian Phenotype Terms

98058 edges
834 edges
20576 edges
Tripartite

ABA – Thalamus
ABA - Striatum dorsal region
ABA - Midbrain
ABA - Medulla
ABA - Retrohippocampal region
ABA - Pallidum
ABA - Olfactory bulb
ABA - Hippocampal formation
ABA - Cerebral cortex
ABA - Hypothalamus

Abnormal developmental patterning
Decreased placenta weight
Abnormal anterior uvea morphology
Abnormal viscerocranium morphology
Increased incidence of induced tumors
Abnormal mitochondrial physiology
Abnormal glycogen homeostasis
Abnormal renal plasma flow rate
Gliosis
Abnormal vascular smooth muscle physiology

NCX NEPCOUNT60 M BXD ILM6v1.1 RankInv
NCX NEPCOUNT45 M BXD ILM6v1.1 RankInv
Multipartite / $K$-partite

Genes

Species

Diseases

Phenotypes

Transcripts
Gulf Coast Analysis

- Residential Retardation Centers
- Mental Health Centers
- Non Mental Health MDs
- Nursing Centers
- Dentists
- Total Birth
- Prenatal Care
- Meat Markets
- Convenience Stores
- Supermarkets
- Fish Markets
- Emergency Relief Services
- Massive/Severe Damage
- Major Damage
- # of Housing Units
- # of Business Establishments
- Educational Services
- Political Associations
- Labor Associations
- Grant Making Associations

# of Business Establishments

Total Arrests

# of Housing Units
Sample Result: 
Putative Prostate Cancer Biomarkers

<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>KLK3</td>
<td>PSA</td>
</tr>
<tr>
<td>ETS1</td>
<td></td>
</tr>
<tr>
<td>MAZR</td>
<td></td>
</tr>
<tr>
<td>KROX</td>
<td></td>
</tr>
<tr>
<td>NFKB</td>
<td></td>
</tr>
</tbody>
</table>

Perform Assorted Forms of Differential Analysis to Identify Network Differences

Select Thresholds, Extract Cliques

Persons Drive Gene-Gene Correlations

Classify Subtypes and Eliminate Outliers
Domain Knowledge

• Lack of grocery stores -> Poor health outcomes
Domain Knowledge

• Lack of grocery stores -> Poor health outcomes
Future Directions

• Modelling large heterogeneous data
• Maximum $k$-partite clique enumeration
  – Vertex Maximum
  – Edge Maximum
References


References


• C. A. Tovey, Tutorial on computational complexity. Interfaces 32, 3, 30-61, 2002.


Acknowledgements

- Mike Langston
- Elissa Chesler
- Erich Baker
- Kai Wang
- Ron Hagan
- Allan Yu
- Gary Rogers
Questions